Major Article.

Genomic analysis of isolates from the UK 2012 pertussis outbreak reveals that vaccine antigen genes are unusually fast evolving.

Short title. Genomics of UK pertussis outbreak.

Katie L. Sealey,[a,c] Simon R. Harris,[b] Norman K. Fry,[c] Laurence D. Hurst,[a] Andrew R. Gorringe,[d] Julian Parkhill,[b] and Andrew Preston[a].

[a]Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, U.K., [b]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, U.K., [c]Public Health England - Respiratory and Vaccine Preventable Bacteria Reference Unit, 61 Colindale Avenue, NW9 5EQ, London, U.K., [d]Public Health England, Porton Down, Salisbury, SP4 0JG, U.K.

Corresponding author: Dr Andrew Preston, Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, U.K. Tel: 44 1225 386318. Email: a.preston@bath.ac.uk

Word count abstract: 149

Word count text: 3500

Conflicts of Interest Statement.

The authors declare that they have no commercial or other association that might pose a conflict of interest.

Financial Support.

This work was supported by a Public Health England PhD Studentship to NKF, ARG, AP and Wellcome Trust (098051) to SRH, JP.

Address for Correspondence.

Dr Andrew Preston, Department of Biology and Biochemistry, University of Bath, Bath, BA2 7AY, U.K. Tel: 44 1225 386318. FAX: 44 1225 386779. Email: a.preston@bath.ac.uk

Abstract

A major outbreak of whooping cough, or pertussis, occurred in 2012 in the U.K, with nearly 10 000 laboratory-confirmed cases and 14 infant deaths attributed to pertussis. A worldwide resurgence of pertussis has been linked to switch to the use of acellular pertussis vaccines and the evolution of *B. pertussis* away from vaccine-mediated immunity. We have conducted genomic analyses of multiple strains from the UK outbreak. We show that the UK outbreak was polyclonal in nature, caused by multiple distinct but closely related strains. Importantly, we demonstrate that acellular vaccine antigen encoding genes are evolving at higher rates than other surface protein encoding genes. This was true even prior to the introduction of pertussis vaccines, but has become more pronounced since the introduction of the current acellular vaccines. The fast evolution of vaccine antigen genes has serious consequences for the ability of current vaccines to continue to control pertussis.

Keywords: Pertussis, genomics, evolution, vaccine

Introduction

Whooping cough, or pertussis, is caused primarily by the bacterium *Bordetella pertussis*. In England and Wales a total of 9,711 laboratory-confirmed cases were recorded in 2012, leading to fourteen deaths in infants under 3 months of age. This was much greater than the previous recent 'peak' year in 2008, in which 902 cases were reported despite levels of vaccine coverage and diagnostic methods not changing during this period [1, 2]. Similar outbreaks have been reported across the globe [3], contributing to the consensus that pertussis is a resurgent disease that might be no longer effectively controlled by current vaccination programmes.

Resurgence has been linked to increased surveillance, better diagnostic techniques, incomplete vaccination of populations but primarily to switching from the use of whole cell (WCV) to acellular (ACV) pertussis vaccines that contain between 1 and 5 purified *B. pertussis* protein antigens: pertussis toxin (Ptx), filamentous haemagglutinin (FHA), pertactin (Prn) and fimbrial types 2 and 3 (Fim2/Fim3). In the UK, a five antigen ACV has been used. ACV induced immunity appears shorter lived than that induced by WCVs, possibly resulting in an expanded pool of carriers, particularly adolescents, and decreased herd immunity [4, 5]. In addition, studies using an infant baboon model revealed that while ACVs protect the individual from disease symptoms, they are less able to prevent colonisation of, and transmission from, the vaccinee compared to WCVs. Increased transmission of *B. pertussis* in populations using ACVs compared to those using WCVs is proposed to contribute to resurgence [6]. Finally, it has been proposed that vaccine escape mutants are arising, as ACV-induced immunity is focused on just a few antigens, and changes in these antigens might result in strains that are less well recognised by this immunity [7].

The frequency of different alleles of vaccine antigen genes among strains has changed over time [8-11]. The most common allelic profile among currently circulating strains (*ptxA1-ptxP3, prn2, fim3-2, fim2-1*) is different to that of strains used for vaccine manufacture [12, 13] and isolates that do not express Prn are increasingly common [14-16]. *PtxP* refers to alleles of the *ptx* promoter. *PtxP3* is now dominant worldwide [17] and some studies suggest that *ptxP3* strains may have increased virulence compared to *ptxP1* strains [18].

The study of genetic changes in *B. pertussis* over time was hindered by the high levels of homogeneity among *B. pertussis* and the lack of fine-resolution tools. Thus recently the genome sequences of a large panel of *B. pertussis* strains collected from around the world and across many decades were generated and analysed [19]. This provided detailed information about the population structure and evolution of *B. pertussis* revealing significant genetic changes among strains over the last 50 years. A lack of geographical clustering of strains suggested rapid strain flow between countries. However, this panel of strains did not contain isolates collected more recently than 2008, except for 3 isolates from the Netherlands collected in 2009 and 2010, and did not intensively sample a specific outbreak meaning that the genetic make-up of such events is largely unknown. Here we analyse a large panel of UK strains with a focus on strains from the recent UK outbreak with the aims of understanding the clonal structure of the outbreak and determining if there is evidence for vaccine-mediated immunity driving the evolution of these strains.

Methods

Accession Numbers

Genome sequence data has been deposited in the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/), Supplementary Table 1.

B. pertussis Strains

100 *B. pertusiss* isolates were obtained from the National Reference Laboratory, Respiratory and Vaccine Preventable Bacteria Reference Unit at Public Health England (Supplementary Table 1). Five strains were collected between 1920-1956

(we define this as the 'pre-vaccine' era), six strains collected between 1957-2000 (WCV era) and 89 strains were collected between 2000-2012 (ACV era). Serotyping was performed using sera specific for antigens 1, 2, and 3 (89/596, 89/598, and 89/600, respectively; National Institute for Biological Standards and Controls, Potters Bar, United Kingdom) as previously described [12]. Tohama I (accession number BX470248), a strain isolated in Japan in 1954, is the most widely studied strain, provides the reference genome sequence of *B. pertussis* [20], and is one of the strains used to produce ACVs used in the UK. *B. pertussis* isolates were grown on charcoal agar for 72 hours at 37°C.

DNA Preparation

Genomic DNA extraction was performed using the Qiagen DNA prep kit according to the manufacturer's instructions.

DNA Sequencing and Single Nucleotide Polymorphism (SNP) Identification

Twenty four isolates were sequenced previously [19]. For the remainder, multiplex libraries, with fragment sizes between 300 and 500bp, were prepared as previously described [21] with modifications [22]. Reads for each isolate were aligned to the Tohama I reference genome using SMALT version 0.7.4 (http://www.sanger.ac.uk/resources/software/smalt/). Base calls were made as previously described [21], using a combination of samtools, mpileup and bcftools [23], allowing SNPs, and small insertions and deletions relative to Tohama I to be identified. Five strains produced poor quality sequence and were excluded from the analysis, resulting in 95 strains being taken forward for analysis.

Phylogenetic Analysis

Maximum likelihood phylogenetic analysis was carried out on variable sites from across the whole genomes using RAxML under a GTR evolutionary model and a gamma correction for among site rate heterogeneity [24]. 100 random bootstrap replicates were run to provide support for relationships identified in the tree.

Analysis of SNP Densities

SNPs were reconstructed on to the phylogenetic tree using parsimony. SNP densities (SNP/bp) within vaccine antigen genes (9 genes: *fhaB*, *prn*, *fim2*, *fim3*, *ptxA-E*) or 'cell surface' functional category genes (591 genes, as categorised previously [20]) were calculated by counting the number of SNPs per bp of each gene. The difference between the mean per gene SNP densities of vaccine antigen genes and cell surface genes was calculated. The significance of this difference was calculated using a non-parametric Monte Carlo simulation. In our randomizations of all the data, preserving relative sample sizes, it was observed how often a difference as large, or greater than the difference above, by repeated randomly resampling two samples of the same size as above. Under this protocol, if n is the number of observations that have greater than or equal to the observed difference in SNP density and m is the number of simulations (in this case, 10 000), then $P = (n+1)/(m+1)$ is the unbiased estimator.

We performed a similar procedure to compare SNP densities in vaccine antigen genes between eras. To account for differences in SNP densities between strains from the different eras, the SNP densities of the vaccine antigen genes were normalised by the SNP densities of all the genes considered (vaccine antigen and surface protein encoding genes). A non-parametric Monte Carlo simulation compared the normalised

164 SNP densities in the ACV antigen genes in ACV-era strains with pre-ACV era strains,
165 with P determined as above.
166
167 Allele Typing
168 The different alleles of *prn*, *ptxA*, *ptxP*, *fim3* and *fim2* genes have been previously
169 described [11] and were used to identify allele types from DNA sequence.
170
171 Analysis of *prn* from UK50
172 The *prn* locus was amplified from UK50 by PCR using primers 5′-
173 CCGCTGATTCGCCACAAG-3′ and 5′-GTGCGGTACTTGCCCTTG-3′. PCR
174 products were cloned using the Gateway system (Invitrogen, Paisley, U.K.) and
175 sequenced by Eurofins Genomics (Ebersberg, Germany) utilising standard M13
176 forward and reverse primers and internal primers 5′-GCGCACGCCTGTCCAAAG-3′
177 and 5′-TAGCGAGCCAGCACGTAG-3′.
178
179 Analysis of Differences in DNA Content Among Strains
180 To detect gene loss from strains, compared to the DNA content of Tohama I, coverage
181 plots generated using the paired end reads mapped to this reference genome were used
182 to create a heat map. DNA sequence contigs that did not map to Tohama I were
183 analysed using Blastn and Blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi).
184
185 Results
186 Phylogeny of UK Strains from 1920-2012
187 Phylogenetic analysis based on SNPs across the whole genome sequences was
188 performed to understand the evolutionary relationships between the UK strains

189 analysed (Fig.1, Suppl. Table 1). Strains isolated during 1920-1982 form a cluster and
190 are generally separated from strains from 2008-2012. The most distinct clustering
191 separates strains carrying the *ptxP1* allele from those carrying the *ptxP3* allele, which,
192 as found elsewhere, is the predominant *ptxP* type among recent strains. This
193 phylogenetic analysis was extended to place the UK strains in the global phylogenetic
194 tree described elsewhere [19] (Fig. 2). This reveals that the UK *ptxP3* strains separate
195 into two clusters, distinguished by the presence of the *fim3-2* allele. The UK outbreak
196 strains largely cluster with strains isolated mainly during the early 2000s from a
197 variety of geographical areas including North America, Europe and Australia.
198
199 Vaccine antigen allele profiles
200 Previously, *ptxP3-ptxA1-prn2-fim*3-2 was defined as the dominant allele type
201 circulating in the UK and other countries [17]. Typing of alleles among the outbreak
202 strains reveal no recent change in this profile (Table 1). Numerous isolates deficient
203 for the production of Prn have been reported in other countries, and a number of
204 different mutations in *prn* responsible for this phenotype identified [14-16]. It has
205 been suggested that loss of Prn expression has been selected by vaccine-mediated
206 immunity pressure. Interestingly, just a single UK strain, UK50, was mutated for *prn*.
207 This was identified by a lack of sequence reads mapping to a region of *prn.* The *prn*
208 locus was amplified by PCR from this strain and the resulting product sequenced
209 using Sanger sequencing. This identified that a recombination event between two
210 copies of IS1663 has resulted in a deletion/insertion mutation in which the 5′ 1326 bp
211 of the *prn* coding sequence has been deleted. Aberrant mapping was not observed for
212 any other UK strain. In other countries, a common *prn* mutation arose from insertion
213 of IS481 into *prn.* We identified paired-end reads in which one read mapped within

IS481 but the other did not and thus derives from the region flanking IS481. Mapping these reads to the reference genome identified the position of the copies of IS481 within each query strain. No IS481 insertions into *prn* were identified among UK strains. It is not clear why so few Prn-deficient strains, compared to other countries experiencing pertussis outbreaks, have been identified in the UK.

SNPs Specific to *ptxP3* Strains

PtxP3 strains are the predominate type in current circulation and appear to have different infection biologies compared to *ptxP1* strains. The *ptxP3* SNP itself appears to be both a direct cause and a marker for other genetic variations that contribute to this difference [18]. To investigate the genetic traits of UK *ptxP3* strains, SNPs specific to this lineage were identified. In total, 22 such SNPs were identified (Table 2). Ten were intergenic, seven of which were in the direct repeat region of IS elements which are present in multiple copies in the *B. pertussis* genome. It is not clear if these particular IS elements are functional. Twelve SNPs were in coding regions. Of those, seven were non-synonomous mutations (NSM) and 5 were synonymous (SM). The 7 NSM were in genes within the "transport and binding proteins", "pseudogene", "conserved hypothetical", "virulence-associated", "unknown" and "regulation" functional categories as defined previously [20]. All of these SNPs were also identified among the global panel of *B. pertussis* strains [19]. However, of the 22 SNPs identified here as being *ptxP3*-specific, only 10 were identified as being *ptxP3*-type-specific in the previous study (Table 2), the other 12 SNPs were also identified among non-*ptxP3* strains globally.

SNP Rates are high in Vaccine Antigen Encoding Genes

Previously, it was identified that genes in the 'cell surface' functional category had higher SNP densities than the *B. pertussis* chromosomal average [19]. However, ACV vaccine-mediated immunity is exerting selective pressure primarily on the proteins used in these vaccines and might be driving their evolution. To explore this, the SNP density (SNPs per bp) for the 9 ACV antigen genes (Ptx comprises five different proteins) and for the other 591 genes comprising the 'cell surface' category was calculated for all strains within each vaccine era and compared. Secondly, it was investigated if the SNP rate in ACV genes had increased since the introduction of ACVs.

The difference in mean SNP density across genes within the two samples (mean SNP density in vaccine antigen genes minus mean SNP density in cell surface genes) was calculated. A non-parametric Monte Carlo simulation was used to assess the significance of this difference by determining how often a difference as large or larger than this was derived by randomly resampling two samples the same size as above, from the pool of vaccine antigen and cell surface genes. This revealed that in each era, vaccine antigen encoding genes had significantly higher SNP densities than other cell surface genes ($P<0.05$, Table 3), with the difference being greatest among ACV-era strains. This suggests that the vaccine antigen genes are faster evolving than other surface protein encoding genes, and that they were also faster evolving even prior to the introduction of widespread vaccination.

To compare SNP densities in vaccine antigen genes between eras, SNP densities within each era were normalised by dividing by the mean SNP rate across all of the genes concerned (ACV antigens and cell surface). In comparison to the prior analysis

this has less power owing to the much smaller sample of ACV genes compared with total cell surface genes. Although the normalised SNP density in ACV-era strains was greater than in pre-ACV era strains, the difference was not statistically significant, P=0.160. However, the number of pre-ACV strains in this analysis was small. Thus, the same analyses were repeated using SNP data from the global collection of strains, for which the year of isolation was known [19], and incorporating the UK strains sequenced here, Table 3. Again, a significantly greater SNP frequency was found in ACV antigen genes than other cell surface genes, in all of the three eras. This time, there was also a significantly higher SNP frequency in ACV genes among ACV era strains compared to pre-ACV era strains (P=0.0177) suggesting that the relative SNP density in ACV antigen genes has increased since the introduction of ACVs. These results suggest that ACV genes are intrinsically fast evolving and provide some support for the hypothesis that they are even faster evolving since the introduction of ACVs.

The more rapid evolution in the ACV antigen genes could be due to either a higher underlying mutation rate or different selection at the protein level. The different selection could be positive selection or weaker purifying selection. To distinguish between these two possibilities, SNPs were split into SM and NSM. High NSM but not SM rates would suggest altered protein-level selection. A higher rate of synonymous evolution (with possibly a weak non-synonymous effect) would suggest higher mutation rates. Interpretation here is difficult owing to well-described but incompletely understood correlation between synonymous and non-synonymous rates.

Among WCV- and ACV-era global strains, but not pre-vaccine era strains, the SM frequency was significantly higher in ACV antigen genes compared to other cell surface genes, Table 4. When comparing ACV-era to pre-ACV era strains, the SM frequency in ACV antigen genes was significantly higher (P=0.004). NSMs also occurred at significantly greater frequency in ACV antigen genes compared to other cell surface genes (Table 4). The magnitude of this effect is greater than that seen for SMs suggesting the higher evolutionary rate of ACV antigen genes compared to cell surface proteins is largely owing to protein-level selection on the antigens. Evidence for a strong recent increase is less clear-cut. When comparing strains from the ACV-era to pre-ACV era strains, the NSM frequency in ACV antigen genes was on the edge of significance (P=0.051). Overall, our results provide support for the hypothesis that the genes encoding antigens chosen for ACVs are intrinsically fast evolving, in part owing to selection on their antigenic products. We cannot discount the possibility that in the ACV-era there has been an increase in the mutation rate (but see also below).

Regions of Difference

Deletions have been a major feature of *B. pertussis* evolution and appear to be on-going [20, 25]. Compared to the Tohama I reference genome, most of the major deletions observed among the strains analysed here had been identified previously [25]. Numerous small deletions were found in only a few, or just one isolate, suggesting that deletion of DNA is common among *B. pertussis* strains. Interestingly, some deletions appeared specific to the UK *ptxP3* strains but no deletions specific to outbreak isolates were detected (Suppl. Fig 1).

313 Regions from individual strains that were not present in the Tohama I reference
314 genome were investigated by BLAST analyses. These regions were also found within
315 other *B. pertussis* genomes (BP18323 and CS), or in *B. bronchiseptica* RB50, similar
316 to that reported in other studies [26]. Thus there were no novel insertions or gene
317 acquisition among outbreak isolates.
318
319 Discussion
320 The resurgence of pertussis in countries with high levels of vaccination has caused
321 widespread concern. Among other factors, *B. pertussis* evolution away from efficient
322 control by vaccine-induced immunity has been proposed as a contributor to this.
323 Recently, whole genome sequencing was used to define global genetic variability
324 among *B. pertussis* isolates and this identified genetic changes in the *B. pertussis*
325 population over time [19].
326
327 Here we have analysed in detail the genomes of UK *B. pertussis* isolates with
328 emphasis on strains from the 2012 outbreak. For the first time we show that many
329 genetically distinct *B. pertussis* strains contributed to this outbreak and importantly,
330 that it was not due to the emergence of a novel, hypervirulent clone or expansion of
331 an individual lineage. Furthermore, outbreak strains were genetically very similar to
332 those circulating during periods when the incidence of pertussis was low.
333
334 The *ptxP3* type is the dominant clone world-wide and UK outbreak strains are also
335 predominantly of this type. Analysis of global isolates identified just 19 SNPs as
336 being *ptxP3*–specific [19]. Here, 22 SNPs distinguished *ptxP3* from *ptxP1* strains.
337 However, just 10 of these were common to both sets of *ptxP3*-specific SNPs. If *ptxP3*

338 strains have increased fitness or virulence compared to older isolates, our analysis
339 suggests that very few SNPs are responsible for this, or that particular combinations
340 of SNPs are important, only some of which are *ptxP3*-specific. Overall, these data
341 argue against large-scale genetic changes being behind the recent resurgence in
342 pertussis.
343
344 Changes in alleles of the genes encoding vaccine antigens have been well documented
345 (for example, [27]) and supports the hypothesis that selection pressure from ACV
346 induced immunity is a driver of *B. pertussis* evolution. However, definitive studies to
347 demonstrate that allelic variation enhances evasion of vaccine-mediated immunity are
348 lacking and particularly difficult to perform given the inability to conduct studies with
349 human hosts and that studies using animal models struggle to detect subtle changes
350 and will not include population level effects that are certainly important for selection
351 of variants among *B. pertussis* worldwide. Here we provide compelling evidence that
352 genes encoding ACV antigens are evolving more rapidly than other cell surface genes
353 (which we consider the most suitable comparator group), containing a significantly
354 higher frequency of SNPs in each of the vaccine eras. Interestingly, this was true even
355 in the pre-vaccine era. It is likely that even in the absence of vaccination, the natural
356 immune response to these antigens creates selective pressure, particularly for a
357 pathogen that is restricted to the human respiratory tract. Of particular importance is
358 that we calculated that ACV antigen gene evolution rates have increased significantly
359 since the introduction of ACVs, the first demonstration of this effect. This might
360 suggest that the use of ACVs has increased selection pressure on ACV antigens,
361 selecting for ACV antigen gene variants. However, we also calculated that while the
362 frequency of SM in ACV antigen genes was significantly higher in ACV era strains

363 compared to older strains, the frequency of NSM was on the edge of significance
364 (P=0.051). In turn this suggests that selection pressure from vaccine-mediated
365 immunity is not the sole driving force for ACV antigen gene variation. A different
366 interpretation is that the mutation rate of ACV antigen genes has increased since the
367 introduction of ACVs. If synonymous sites are under weak purifying selection (i.e.
368 not perfectly neutral), then there is a lag between a SNP arising and its elimination by
369 this selection, resulting in an excess of SNPs in the modern era. However, normalising
370 ACV gene SNP rates by the SNP rates for all genes within the era largely eliminates
371 this effect (i.e. SMs in cell surface genes should be equally over-represented in the
372 modern era). However, if SMs in ACV genes and cell surface genes are under
373 different intensities of purifying selection, then our result could be found.
374
375 Either way, the more rapid evolution at the protein level (as determined by NSM) of
376 ACV proteins compared to other cell surface proteins, across all eras suggests that
377 strains will become increasingly mismatched to those used for vaccine production and
378 this could lead to decreased vaccine efficacy over time. The ACV antigens were
379 chosen on the basis of their immunogenicity but it could be that this property has
380 driven the relatively high evolution rates of the genes encoding these antigens. Our
381 results raise fresh concerns over the ability of current acellular pertussis vaccines to
382 continue to control disease.
383
384 Funding
385 This work was supported by a Public Health England PhD Studentship to NKF, ARG,
386 AP and Wellcome Trust (Grant number 098051) to SRH, JP.
387

388 References

389 1. Public Health England. Enhanced Pertussis Surveillance. **2014.** Available at:
390 http://www.hpa.org.uk/webc/HPAwebFile/HPAweb_C/1317133571726. Accessed
391 November 2014.
392 2. Public Health England. Available at
393 https://www.gov.uk/government/publications/whooping-cough-pertussis-statistics.
394 Accessed November 2014.
395 3. Jakinovich A, Sood SK. Pertussis: still a cause of death, seven decades into
396 vaccination. Curr Opin Pediatr **2014**; 26:597-604.
397 4. Rendi-Wagner P, Kundi M, Mikolasek A, Vecsei A, Fruhwirth M, Kollaritsch H.
398 Hospital-based active surveillance of childhood pertussis in Austria from 1996 to
399 2003: estimates of incidence and vaccine effectiveness of whole-cell and acellular
400 vaccine. Vaccine **2006**; 24:5960-5.
401 5. Witt MA, Arias L, Katz PH, Truong ET, Witt DJ. Reduced risk of pertussis among
402 persons ever vaccinated with whole cell pertussis vaccine compared to
403 recipients of acellular pertussis vaccines in a large US cohort. Clin Infect Dis
404 **2013**; 56:1248-54.
405 6. Warfel JM, Zimmerman LI, Merkel TJ. Acellular pertussis vaccines protect
406 against disease but fail to prevent infection and transmission in a nonhuman
407 primate model. Proc Natl Acad Sci U S A **2013**; 111:787-92.
408 7. Poolman JT. Shortcomings of pertussis vaccines: why we need a third
409 generation vaccine. Expert Rev Vaccines **2014**; 13:1159-62.
410 8. Bottero D, Gaillard ME, Basile LA, Fritz M, Hozbor DF. Genotypic and
411 phenotypic characterization of *Bordetella pertussis* strains used in different
412 vaccine formulations in Latin America. J Appl Microbiol **2012**; 112:1266-76.

9. Elomaa A, Advani A, Donnelly D, et al. Population dynamics of *Bordetella pertussis* in Finland and Sweden, neighbouring countries with different vaccination histories. Vaccine **2007**; 25:918-26.

10. Komatsu E, Yamaguchi F, Abe A, Weiss AA, Watanabe M. Synergic effect of genotype changes in pertussis toxin and pertactin on adaptation to an acellular pertussis vaccine in the murine intranasal challenge model. Clin Vaccine Immunol **2010**; 17:807-12.

11. Mooi FR. *Bordetella pertussis* and vaccination: the persistence of a genetically monomorphic pathogen. Infect Genet Evol **2010**; 10:36-49.

12. Litt DJ, Neal SE, Fry NK. Changes in genetic diversity of the *Bordetella pertussis* population in the United Kingdom between 1920 and 2006 reflect vaccination coverage and emergence of a single dominant clonal type. J Clin Microbiol **2009**; 47:680-8.

13. Van Loo IH, Mooi FR. Changes in the Dutch *Bordetella pertussis* population in the first 20 years after the introduction of whole-cell vaccines. Microbiology **2002**; 148:2011-8.

14. Bouchez V, Brun D, Cantinelli T, Dore G, Njamkepo E, Guiso N. First report and detailed characterization of *B. pertussis* isolates not expressing Pertussis Toxin or Pertactin. Vaccine **2009**; 27:6034-41.

15. Lam C, Octavia S, Ricafort L, et al. Rapid increase in pertactin-deficient *Bordetella pertussis* isolates, Australia. Emerg Infect Dis **2014**; 20:626-33.

16. Otsuka N, Han HJ, Toyoizumi-Ajisaka H, et al. Prevalence and genetic characterization of pertactin-deficient *Bordetella pertussis* in Japan. PloS One **2012**; 7:e31985.

437 17. Kallonen T, He Q. *Bordetella pertussis* strain variation and evolution
438 postvaccination. Expert Rev Vaccines **2009**; 8:863-75.
439 18. King AJ, van der Lee S, Mohangoo A, van Gent M, van der Ark A, van de
440 Waterbeemd B. Genome-wide gene expression analysis of *Bordetella pertussis*
441 isolates associated with a resurgence in pertussis: elucidation of factors involved
442 in the increased fitness of epidemic strains. PloS One **2013**; 8:e66150.
443 19. Bart MJ, Harris SR, Advani A, et al. Global population structure and evolution
444 of *Bordetella pertussis* and their relationship with vaccination. MBio **2014**; 5.
445 20. Parkhill J, Sebaihia M, Preston A, et al. Comparative analysis of the genome
446 sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella*
447 *bronchiseptica*. Nat Genet **2003**; 35:32-40.
448 21. Harris SR, Feil EJ, Holden MT, et al. Evolution of MRSA during hospital
449 transmission and intercontinental spread. Science **2010**; 327:469-74.
450 22. Quail MA, Otto TD, Gu Y, et al. Optimal enzymes for amplifying sequencing
451 libraries. Nat Methods **2012**; 9:10-1.
452 23. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools.
453 Bioinformatics **2011**; 27:2156-8.
454 24. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic
455 analyses with thousands of taxa and mixed models. Bioinformatics **2006**;
456 22:2688-90.
457 25. Caro V, Bouchez V, Guiso N. Is the sequenced *Bordetella pertussis* strain
458 Tohama I representative of the species? J Clin Microbiol **2008**; 46:2125-8.
459 26. Kallonen T, Grondahl-Yli-Hannuksela K, Elomaa A, et al. Differences in the
460 genomic content of *Bordetella pertussis* isolates before and after introduction of

461 pertussis vaccines in four European countries. Infect Genet Evol **2011**; 11:2034-
462 42.
463 27. van Gent M, Bart MJ, van der Heide HG, Heuvelman KJ, Mooi FR. Small
464 mutations in *Bordetella pertussis* are associated with selective sweeps. PloS One
465 **2012**; 7:e46407.
466
467
468 Figure Legends.
469 Figure 1. Phylogenetic tree depicting the evolutionary relationships among the UK *B.*
470 *pertussis* isolates studied here. Maximum likelihood (ML) phylogenetic analysis was
471 carried out on variable sites from across the whole genomes using RAxML. Strains
472 are shaded according to their year of isolation and *ptxP* type.
473
474 Figure 2. Phylogenetic relationships of UK strains within a global context. The UK
475 isolates analysed here are indicated.
476
477 Supplemental Table 1.
478 Details of strains analysed in this study.
479
480 Supplemental Figure 1.
481 A heat map of coverage plots of the sequence reads of each UK strain mapped to the
482 Tohama I reference genome. Black regions indicate sequence common to both query
483 and reference genomes, white regions indicate regions of the reference genome that
484 are absent from the query strain genomes.

485 Table 1. Frequency (% of strains tested) of vaccine antigen encoding gene alleles
486 among UK strains.

		Period		
		Prevaccine 1920-1956	**WCV 1957-2000**	**ACV 2001-2012**
No. of strains		5	6	84
ptxP	**1**	100	100	6
	3	0	0	94
ptxA	**1**	20	100	100
	2	80	0	0
***Prn**	**1**	100	84	5
	2	0	0	91
	3	0	16	3
	4	0	0	1
Fim2-	**1**	100	100	100
Fim3-	**1**	100	100	70
	2	0	0	29
	3	0	0	1
****Serotype**	**1**	20	0	0
	1,2	40	50	37
	1,3	20	17	63
	1,2,3	20	33	0

487

*Prn allele type was determined for just 76 ACV era strains due to poor mapping of reads in this region in 8 strains.

** Serotype was not determined for one ACV era strain, thus frequencies are based on 83, not 84, strains in this era.

493 Table 2. SNPs specific to UK *ptxP3* strains.

Location[a]	Type[b]	Mutation[c]	Global *ptxP3*[d]	Details
36857	INT	A:G	Yes	93 bp upstream of BP0032 (encoding a putative transport protein), 156bp upstream of BP0033 (encoding GlyQ-glycyl-tRNA synthetase alpha chain)
617083	INT	T:G	No	within the 5' repeat region of IS481 (BP0611). 31bp upstream of transposase start codon.
617084	INT	C:T	No	within the 5' repeat region of IS481 (BP0611). 32bp upstream of transposase start codon.
1077844	INT	C:T	No	within the 5' repeat region of IS1663 (BP1035). 139bp upstream of transposase start codon.
1170424	INT	A:G	No	within the 5' repeat region of IS481 (BP1114). 31bp upstream of transposase start codon.
1222400	INT	A:C	No	within the 5' repeat region of IS481 (BP1157). 31 bp upstream of transposase start codon.
1635654	INT	T:G	No	within the 5' repeat region of IS481 (BP1557). 31 bp upstream of transposase start codon.
2259917	INT	G:C	No	within the 5' repeat region of IS481 (BP2135). 98 bp upstream of transposase start codon.
3263622	INT	A:C	Yes	193 bp away from BP3062. Putative integral membrane transport protein.
3988168	INT	G:A	Yes	89 nucleotides away from the start codon of

				ptxA. ptxP3allele.
196307	NSM	T:C	Yes	BP0194. Putative transport protein.
299559	NSM	C:T	Yes	BP0292. Pseudogene. Conserved hypothetical protein.
1331840	NSM	G:A	Yes	Pseudogene. BP1261. Hypothetical protein.
1547488	NSM	A:G	No	BP1471. Conserved hypothetical protein.
2374322	NSM	T:C	Yes	BP2249. BscI. Type III secretion apparatus protein.
2651008	NSM	G:A	Yes	BP2502. Hypothetical protein.
3134458	NSM	G:C	No	BP2946. Probable transcriptional regulator.
185405	SM	G:A	No	BP0184. Putative periplasmic protein.
518837	SM	T:C	No	BP0507. Putative membrane protein.
694521	SM	A:G	Yes	BP0678. Putative peptide chain release factor.
3840411	SM	G:A	Yes	BP3630. RpsH. 30S ribosomal protein.
3991376	SM	C:T	No	BP3787. PtxC. Pertussis toxin subunit protein.

[a] Tohama I reference genome coordinates (accession no. BX470248) .

[b] Int: SNP is in an intergenic region. NSM: non-synonymous mutation, SM: synonymous mutation.

[c] e.g. C:T – C to T mutation.

[d] SNP is also defined as *ptxP3*-specific in study of global *B. pertussis* population (1).

Table 3.

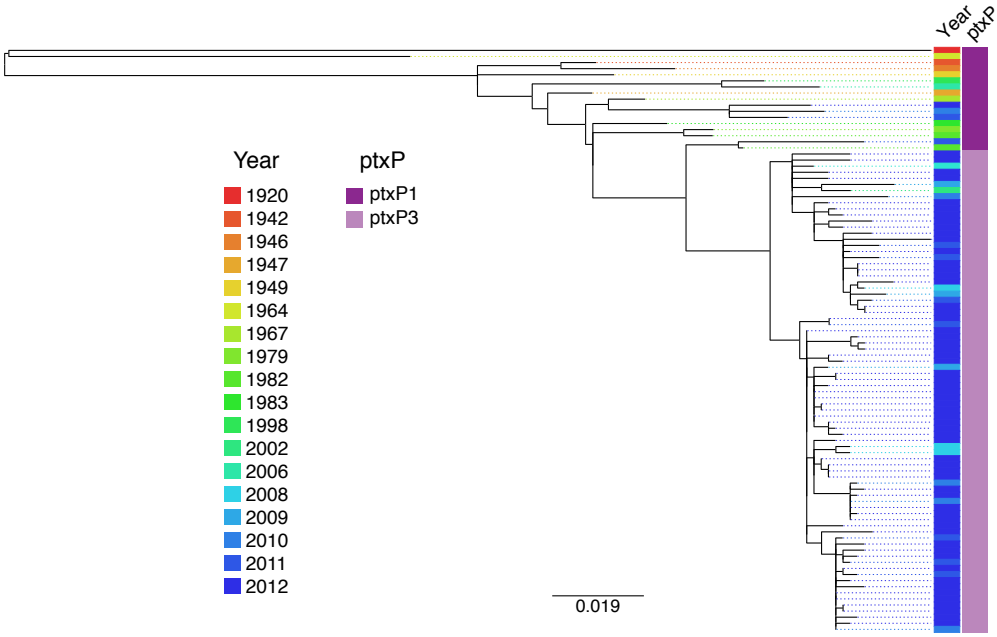
SNP rates in vaccine antigen encoding genes compared to other cell surface genes for the different vaccine eras among UK strains and globally.

Vaccine era. (No. of strains)	**mean SNP/bp vaccine antigen genes**	**mean SNP/bp cell surface genes**	**Diffference (vaccine antigens - cell surface)**	**Difference normalized (Difference/mean SNP density)**	**P (SNP rate vaccine antigens > SNP rate cell surface)**
UK Pre-1920-1956 (5)	3×10^{-4}	7.8×10^{-5}	2.22×10^{-4}	2.72	0.045
UK WCV 1957-2000 (6)	4.75×10^{-4}	5.9×10^{-5}	4.17×10^{-4}	6.40	0.016
UK ACV 2001-2012 (84)	1.73×10^{-3}	1.55×10^{-4}	1.57×10^{-3}	8.82	0.0004
Global Pre-1920-1956 (19)	1.45×10^{-3}	5.85×10^{-4}	8.62×10^{-4}	1.44	0.012
Global WCV 1957-2000 (204)	2.62×10^{-3}	1.01×10^{-3}	1.61×10^{-3}	1.56	0.002
Global ACV 2001-2012 (188)	2.91×10^{-3}	4.23×10^{-4}	2.49×10^{-3}	5.41	0.0001

505 Table 4. Synonomous (SM) and non-synonomous (NSM) mutation rates in vaccine
506 antigen genes compared to other cell surface genes among strains isolated during the
507 different vaccine eras.
508

Vaccine era. (No. of strains)	mean SNP/bp vaccine antigen genes	mean SNP/bp cell surface genes	Diffference (vaccine antigens - cell surface)	Difference normalized (Difference/mean SNP density)	P (SNP rate vaccine antigens > SNP rate cell surface)
SM Global Pre-1920-1956 (19)	1.32×10^{-4}	2.4×10^{-4}	-1.07×10^{-4}	-0.45	0.627
SM Global WCV 1957-2000 (204)	9.66×10^{-4}	4.23×10^{-4}	5.43×10^{-4}	1.26	0.045
SM Global ACV 2001-2012 (188)	9.68×10^{-4}	1.76×10^{-4}	7.92×10^{-4}	4.20	0.011
NSM Global Pre-1920-1956 (19)	1.18×10^{-3}	3.40×10^{-4}	8.38×10^{-4}	2.38	0.006
NSM Global WCV 1957-2000 (204)	1.96×10^{-3}	5.83×10^{-4}	1.37×10^{-3}	2.28	0.002
NSM Global ACV 2001-2012 (188)	1.95×10^{-3}	2.38×10^{-4}	1.71×10^{-3}	6.48	0.0002

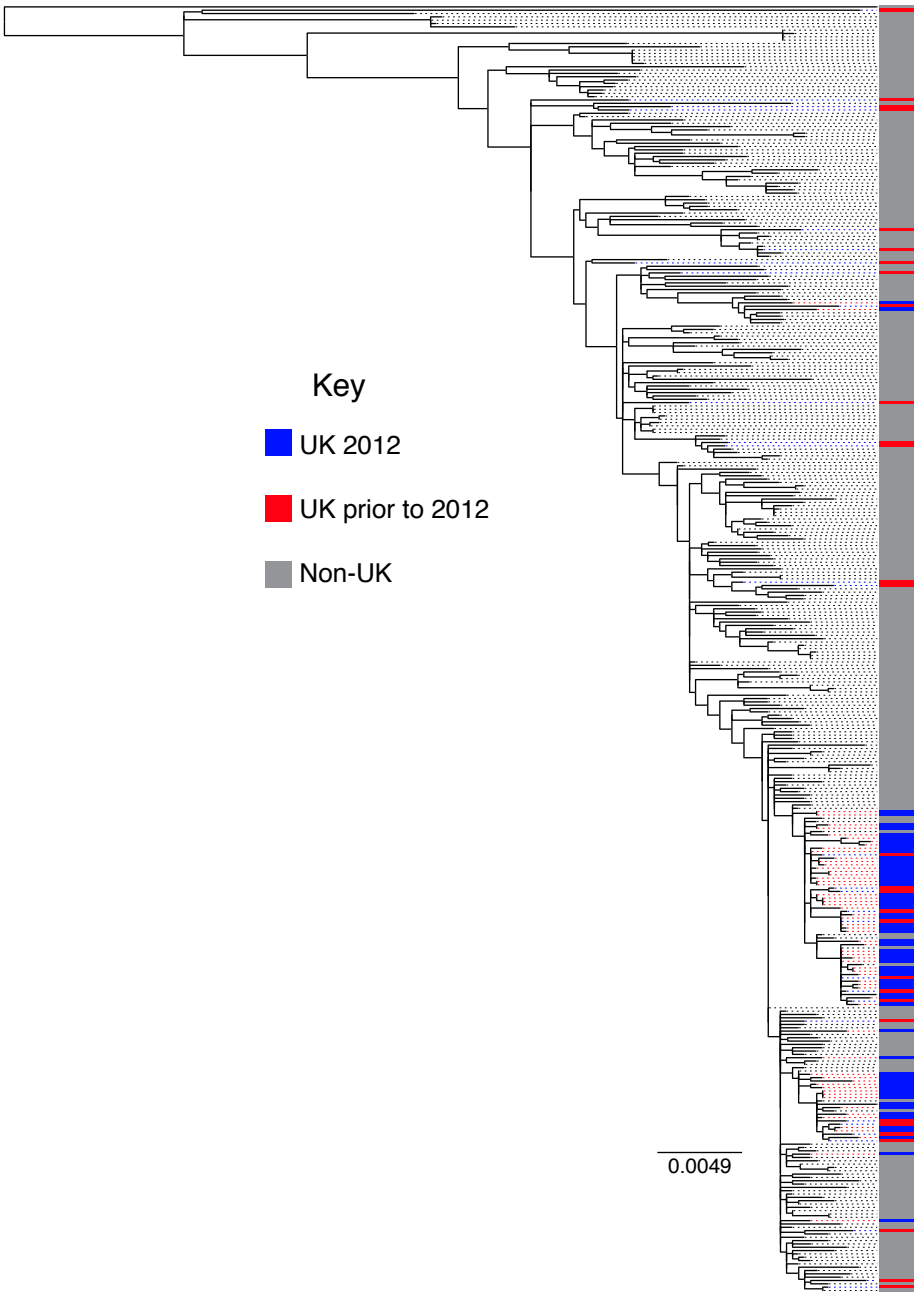
509
510


Year
ptxP
Year
1920
1942
1946
1947
1949
1964
1967
1979
1982
1983
1998
2002
2006
2008
2009
2010
2011
2012
ptxP
ptxP1
ptxP3
0.019

511

512

513



514